APPENDIX B

Estimation and Inference in Econometrics

RUSSELL DAVIDSON
JAMES G. MACKINNON

New York Oxford
OXFORD UNIVERSITY PRESS
1993

If u were asymptotically uncorrelated with X, this quantity would just be σ^2. Instead, it is *smaller* than σ^2. Thus using least squares makes the model fit too well. Because least squares minimizes the distance between y and $\mathcal{S}(X)$, part of the variation in y that is really due to variation in the error terms u has incorrectly been attributed to variation in the regressors.

Unfortunately, there are many situations in econometrics in which the error terms cannot be expected to be orthogonal to the X matrix. We will discuss two of them, the cases of errors in variables and simultaneous equations bias, in Sections 7.2 and 7.3. The most general technique for handling such situations is the method of **instrumental variables**, or **IV** for short. This technique, proposed originally by Reiersøl (1941) and further developed by Durbin (1954) and Sargan (1958), among many others, is very powerful and very general. Numerous variants of it appear in many branches of econometrics. These include **two-stage least squares** (Section 7.5), **three-stage least squares** (Chapter 18), and the **generalized method of moments** (Chapter 17).

The plan of the chapter is as follows. In the next section, we discuss the very common problem of errors in variables, for which the method of instrumental variables was originally proposed as a solution. Then, in Section 7.3, we provide an introduction to the linear simultaneous equations model and show that OLS is biased when applied to one equation of such a model. In Section 7.4, we introduce the method of instrumental variables in the context of a linear regression equation and discuss many of its properties. In the following section, we discuss two-stage least squares, which is really just another name for the IV estimator of the parameters of a linear regression model. In Section 7.6, we show how the IV method may be used to estimate nonlinear regression models. In Section 7.7, we generalize the Gauss-Newton regression to the IV case and discuss how to test hypotheses about the coefficients of regression models when they have been estimated by IV. In Section 7.8, we discuss the issue of identification in regression models estimated by IV. Finally, in Section 7.9, we consider a class of tests called Durbin-Wu-Hausman tests, which may be used to decide whether or not it is necessary to employ instrumental variables.

7.2 ERRORS IN VARIABLES

Almost all economic variables are measured with error. This is true to a greater or lesser extent of all macroeconomic time series and is especially true of survey data and many other cross-section data sets. Unfortunately, the statistical consequences of errors in explanatory variables are severe, since explanatory variables that are measured with error are necessarily correlated with the error terms. When this occurs, the problem is said to be one of **errors in variables**. We will illustrate the problem of errors in variables with a simple example.

Suppose, for simplicity, that the DGP is

$$y = \alpha_0 + \beta_0 x + u, \quad u \sim \text{IID}(0, \sigma_0^2 \mathbf{I}), \tag{7.04}$$

where x is a vector that is observed with error. We actually observe x^*, which is related to x by

$$x^* = x + v, \quad v \sim \text{IID}(0, \omega^2 \mathbf{I}).$$

The vector v is a vector of measurement errors, which are assumed (possibly unrealistically) to have the i.i.d. property and to be independent of x and u. Substituting $x^* - v$ for x in (7.04), the DGP becomes

$$y = \alpha_0 + \beta_0 x^* - \beta_0 v + u.$$

Thus the equation we can actually estimate is

$$y = \alpha + \beta x^* + u^*, \tag{7.05}$$

where $u^* \equiv u - \beta_0 v$. It is clear that u^* is not independent of x^*. In fact

$$E(x^{*\top} u^*) = E\big((x + v)^\top (u - \beta_0 v)\big) = -\beta_0 E(v^\top v) = -n\beta_0 \omega^2,$$

where, as usual, n is the sample size. If we assume for concreteness that $\beta_0 > 0$, the error term u^* is negatively correlated with the regressor x^*. This negative correlation means that least squares estimates of β will be biased and inconsistent, as will least squares estimates of α unless x^* happens to have mean zero. Note that the inconsistency of $\hat{\beta}$ is a problem only if we care about the parameter β. If, on the contrary, we were simply interested in finding the mean of y conditional on x^*, estimating equation (7.05) by least squares is precisely what we would want to do.

There are many ways to deal with the problem of errors in variables, the method of instrumental variables being only one of them. In the above example, it is clear that if we knew ω^2, we could say something about the bias of $\hat{\beta}$ and hence derive a better estimate. This observation has led to various alternative approaches to the errors in variables problem: See Frisch (1934), Klepper and Leamer (1984), Hausman and Watson (1985), and Leamer (1987), among others.

7.3 SIMULTANEOUS EQUATIONS

The reason most commonly cited in applied econometric work for explanatory variables to be correlated with error terms is that the former are determined endogenously, rather than being exogenous or predetermined. A variable that is **predetermined** at time t is one that was determined, possibly endogenously, at some earlier time period. The simplest example is a lagged dependent variable. A detailed discussion of exogeneity and predeterminedness may be found in Section 18.2. Models in which two or more endogenous variables are

11.2 TESTS FOR EQUALITY OF TWO PARAMETER VECTORS

A classic problem in econometrics is determining whether the coefficients of a regression model (usually a linear one) are the same in two (or sometimes more than two) separate subsamples. In the case of time-series data, the subsamples would normally correspond to different time periods, and these tests are then often referred to as tests for **structural change**. Sometimes we may be interested in testing whether the coefficients are the same in two or more different time periods simply as a way of testing whether the model is specified correctly. In such cases, time-series data sets may be divided into earlier and later periods in a fairly arbitrary way for purposes of testing. This is legitimate, but such tests are more interesting when there is reason to believe that the subsamples correspond to different economic environments, such as different exchange-rate or policy regimes.[1] In the case of cross-section data, arbitrary division almost never makes sense; instead, the subsamples might correspond to such potentially different groups of observations as large firms and small firms, rich countries and poor countries, or men and women. In these cases, the results of the test are often of interest for their own sake. For example, a labor economist might be interested in testing whether the earnings functions of men and women or of two different ethnic groups are the same.[2]

The classic treatment of this problem has deep roots in the statistical literature on the analysis of variance (Scheffé, 1959). An early and very influential paper in econometrics is G. C. Chow (1960), and as a result the standard F test for the equality of two sets of coefficients in linear regression models is commonly referred to by economists as the **Chow test**. Fisher (1970) provides a neater exposition of the classic Chow test procedure. Dufour (1982) provides a more geometrical exposition and generalizes the test to handle any number of subsamples, some of which may have fewer observations than there are regressors.

The standard way of posing the problem is to partition the data into two parts, the n–vector $\boldsymbol{y}$ of observations on the dependent variable being divided into two vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, of lengths n_1 and n_2, respectively, and the $n \times k$ matrix $\boldsymbol{X}$ of observations on the regressors being divided into two matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, of dimensions $n_1 \times k$ and $n_2 \times k$, respectively. This partitioning may of course require that the data be reordered. Thus the maintained hypothesis

[1] When there is no reason to expect parameters to have changed at any particular point in time, it may make sense to use a procedure that does not specify such a point. Examples include the CUSUM and CUSUM of squares procedures of Brown, Durbin, and Evans (1975).

[2] An earnings function relates earnings to a number of right-hand side variables, such as age, education, and experience. As examples of the use of F tests for the equality of two sets of coefficients in this context, see Oaxaca (1973, 1974).

MODERN ELEMENTARY STATISTICS

Seventh Edition

John E. Freund
Arizona State University

PRENTICE-HALL / Englewood Cliffs, New Jersey 07632

Since "significant" is often used interchangeably with "meaningful" or "important" in everyday language, it must be understood that we are using it here as a technical term. If a result is statistically significant, this does not mean that it is necessarily of any great importance, or that it is of any practical value. Suppose, for instance, that the social scientist of our example takes her sample and gets $\bar{x} = 1.525$. According to the criterion on page 297 this result is statistically significant—the difference between $\bar{x} = 1.525$ and $\mu = 1.4$ is too big to be attributed to chance—but then nobody may care. Even an insurance company which ought to be interested in such a result may well feel that it is not worth bothering about.

Returning to the airport parking example, we could convert the criterion on page 290 into that of a significance test by writing

> **Reject the hypothesis $\mu = 42.5$ minutes (and accept the alternative $\mu \neq 42.5$ minutes) if the mean of the 50 sample values is less than 40.5 minutes or greater than 44.5 minutes; reserve judgment if the mean falls anywhere from 40.5 to 44.5 minutes.**

So far as the rejection of the null hypothesis is concerned, the criterion has remained unchanged and the probability of a Type I error is still 0.06. However, so far as its acceptance is concerned, the members of the planning commission are now playing it safe by reserving judgment.

Reserving judgment in a significance test is similar to what happens in court proceedings where the prosecution does not have sufficient evidence to get a conviction, but where it would be going too far to say that the defendent definitely did not commit the crime. In general, whether one can afford the luxury of reserving judgment in any given situation depends entirely on the nature of the situation. If a decision must be reached one way or the other, there is no way of avoiding the risk of committing a Type II error.

Since most of the remainder of this book will be devoted to significance tests—indeed, most statistical problems which are not problems of estimation or prediction deal with tests of this kind—it will help to perform such tests by proceeding systematically as outlined in the following five steps. The first of these may look simple and straightforward, yet it presents the greatest difficulties to most beginners.

1. We formulate the null hypothesis and an appropriate alternative.

In the airport parking example the null hypothesis was $\mu = 42.5$ minutes, and the alternative hypothesis was $\mu \neq 42.5$ minutes (presumably because the planning commission wanted to protect itself against the possibility that $\mu = 42.5$ minutes may be too high or too low). We refer to this kind of alternative as a **two-sided alternative**. In the traffic ticket example the null hypothesis was $\mu = 1.4$, and the alternative hypothesis was $\mu > 1.4$ (because the social scientist suspected that licensed drivers over 65 average more than 1.4 tickets per year). This is called a

A Course in Econometrics

Arthur S. Goldberger

Harvard University Press
Cambridge, Massachusetts
London, England

the event $\{|z_j^\circ| > 1.96\}$. The probability that A occurs depends on what the true value of β_j is. If the true value is β_j°, so that the null hypothesis is true, then the random variable z_j° is identical to the random variable z_j defined in

D3A. $z_j = (b_j - \beta_j)/\sigma_{b_j} \sim \mathcal{N}(0, 1)$.

Consequently, $\Pr(A|\beta_j = \beta_j^\circ) = 0.05$. So the significance level, namely the probability of rejecting the null hypothesis when it is true, is 5%.

Suppose a sample has $|z_j^\circ| > 1.96$. If the null is true, then a low-probability event has occurred. The probability of the event is so low that its occurrence is taken to be evidence against the null; so the decision is to reject the null. Heuristically, the point estimate b_j is so far from the hypothesized parameter value β_j° that it is implausible that b_j has in fact been drawn from a distribution with expected value β_j°. However, finding a sample with $|z_j^\circ| \leq 1.96$ is not surprising when the null is true, so then the decision is to accept the null.

When $|z_j^\circ| > 1.96$, one says that b_j *is significantly different from* β_j° at the 5% level; when $|z_j^\circ| \leq 1.96$, one says that b_j *is not significantly different from* β_j° at the 5% level.

Several lessons are immediate:

- Rejection of the null is not proof that the null is false. After all, there is a nonzero probability of rejecting the null if it is true: $\Pr(|z_j^\circ| > 1.96|\beta_j = \beta_j^\circ) = 0.05$. Loosely speaking, when the null is true, in 5% of the samples drawn from the population, the decision will be "reject the null."
- Acceptance of the null is not proof that the null is true. After all, different null hypotheses would also have been acceptable. Indeed if the null had been $\beta_j = \beta_j^{\circ\circ}$, where $\beta_j^{\circ\circ}$ is any other point that happens to lie in the confidence interval $b_j \pm 1.96\sigma_{b_j}$, it too would have been accepted as a null hypothesis.
- If σ_{b_j} is large, then the 95% confidence interval is wide, and widely diverse null hypotheses about β_j are all acceptable at the 5% level. In that situation, the sample contains little information about the true value of β_j. The LS estimator b_j may well be the best estimator, but it need not be a precise estimator.

The test procedure adapts to handle a null hypothesis about β_j at different significance levels. Further, to test a null hypothesis about a single linear combination of the elements of $\boldsymbol{\beta}$: accept iff the null θ° lies

Here is a final example to show how the solving-out-the-restrictions approach works. Suppose that you have a pair of data sets to which

$$E(\mathbf{y}_1) = \mathbf{X}_1\boldsymbol{\beta}_1, \qquad E(\mathbf{y}_2) = \mathbf{X}_2\boldsymbol{\beta}_2$$

apply, where $\mathbf{y}_1$ is $n_1 \times 1$, $\mathbf{X}_1$ is $n_1 \times k$, $\mathbf{y}_2$ is $n_2 \times 1$, and $\mathbf{X}_2$ is $n_2 \times k$. You want to test the null hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. Assemble the data together as

$$E(\mathbf{y}) = \begin{pmatrix} E(\mathbf{y}_1) \\ E(\mathbf{y}_2) \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1\boldsymbol{\beta}_1 \\ \mathbf{X}_2\boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \mathbf{X}\boldsymbol{\beta},$$

say. If the null $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ ($= \boldsymbol{\beta}^\circ$, say) is true, then

$$E(\mathbf{y}) = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \boldsymbol{\beta}^\circ = \mathbf{X}^\circ\boldsymbol{\beta}^\circ,$$

say. The relevant sums of squared residuals are obtainable from a long regression ($\mathbf{y}$ on the $2k$ columns of $\mathbf{X}$) and a short regression ($\mathbf{y}$ on the k columns of $\mathbf{X}^\circ$). Provided that the CNR model applies to each sample, with the same σ^2, while the two samples are independent, the difference between those sums of squared residuals is the kernel of the appropriate F-statistic. This special case of a standard F-test is sometimes referred to as a "test for structural change," or as a "Chow test." Incidentally, the long-regression sum of squared residuals can be calculated by adding together the sums of squared residuals obtained in separate LS regressions of $\mathbf{y}_1$ on $\mathbf{X}_1$, and $\mathbf{y}_2$ on $\mathbf{X}_2$.

22.3. One-Sided Alternatives

To test the single-parameter null hypothesis $\beta_j = \beta_j^\circ$ against the alternative that $\beta_j \neq \beta_j^\circ$, we have learned to use the t-statistic given by $u_j^\circ = (b_j - \beta_j^\circ)/\hat{\sigma}_{b_j}$, rejecting the null iff $|u_j^\circ| > c$, where $G(c) = 0.975$ with $G(\cdot)$ being the cdf of the $t(n-k)$ distribution.

Now suppose, as occurs in some economic contexts, that the known alternative to $\beta_j = \beta_j^\circ$ is *one-sided*, say $\beta_j > \beta_j^\circ$. A *one-tailed* version of the t-test can be used: reject the null $\beta_j = \beta_j^\circ$ iff $u_j^\circ > c^*$, where $G(c^*) = 0.95$. This variant is sensible. Heuristically, it would be foolish to reject $\beta_j = \beta_j^\circ$ *in favor of* $\beta_j > \beta_j^\circ$ when the sample has $b_j < \beta_j^\circ$. More formally, the one-tailed test has more power than the two-tailed test, for all $\beta_j > \beta_j^\circ$

Divide Eq. (16.7) through by $\Sigma_i(y_i - \bar{y})^2$ to get

$$(16.8) \quad R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}.$$

The measure R^2, which will lie between zero and unity, is called the *coefficient of determination*, or squared multiple correlation coefficient. It measures, one says, the proportion of the variation of y that is accounted for (linearly) by variation in the x_j's; note that the fitted value $\hat{y}_i$ is an exact linear function of the x_{ij}'s. In this sense, R^2 measures the goodness of fit of the regression.

Consider an extreme case:

$$R^2 = 1 \Leftrightarrow \sum_i e_i^2 = 0 \Leftrightarrow \mathbf{e}'\mathbf{e} = 0 \Leftrightarrow \mathbf{e} = \mathbf{0} \Leftrightarrow \mathbf{y} = \mathbf{Xb},$$

in which case the observed y's fall on an exact linear function of the x's. The fit is perfect; all of the variation in y is accounted for by the variation in the x's. At the other extreme:

$$R^2 = 0 \Leftrightarrow \sum_i (\hat{y}_i - \bar{y})^2 = 0 \Leftrightarrow \hat{y}_i = \bar{y} \quad \text{for all } i,$$

in which case the best-fitting line is horizontal, and none of the variation in y is accounted for by variation in the x's.

From our perspective, R^2 has a very modest role in regression analysis, being a measure of the goodness of fit of a sample LS linear regression in a body of data. Nothing in the CR model requires that R^2 be high. Hence a high R^2 is not evidence in favor of the model, and a low R^2 is not evidence against it. Nevertheless, in empirical research reports, one often reads statements to the effect that "I have a high R^2, so my theory is good," or "My R^2 is higher than yours, so my theory is better than yours."

In fact the most important thing about R^2 is that it is not important in the CR model. The CR model is concerned with parameters in a population, not with goodness of fit in the sample. In Section 6.6 we did introduce the population coefficient of determination ρ^2, as a measure of strength of a relation in the population. But that measure will not be invariant when we sample selectively, as in the CR model, because it depends upon the marginal distribution of the explanatory variables. If one insists on a measure of predictive success (or rather failure), then $\hat{\sigma}^2$ might suffice: after all, the parameter σ^2 is the expected squared

forecast error that would result if the population CEF were used as the predictor. Alternatively, the squared standard error of forecast (Section 16.3) at relevant values of $\mathbf{x}$ may be informative.

Some further remarks on the coefficient of determination follow.

• One should not calculate R^2 when $\bar{e} \neq 0$, for then the equivalence of the two versions of R^2 in Eq. (16.8) breaks down, and neither of them is bounded between 0 and 1. What guarantees that $\bar{e} = 0$? The only guarantee can come from the FOC's $\mathbf{X}'\mathbf{e} = \mathbf{0}$. It is customary to allow for an intercept in the regression, that is, to have, as one of the columns of $\mathbf{X}$, the $n \times 1$ vector $\mathbf{s} = (1, 1, \ldots, 1)'$. We refer to this $\mathbf{s}$ as the *summer vector*, because multiplying $\mathbf{s}'$ into any vector will sum up the elements in the latter. If $\mathbf{s}$ is one of the columns in $\mathbf{X}$, then $\mathbf{s}'\mathbf{e} = 0$ is one of the FOC's, so $\bar{e} = 0$. The same conclusion follows if there is a linear combination of the columns of $\mathbf{X}$ that equals the summer vector. Also if $\mathbf{y}$ and $\mathbf{x}_2, \ldots, \mathbf{x}_k$ all have zero column means in the sample, then $\bar{e} = 0$. But otherwise a zero mean residual is sheer coincidence.

• We can always find an $\mathbf{X}$ that makes $R^2 = 1$: take any n linearly independent $n \times 1$ vectors to form the $\mathbf{X}$ matrix. Because such a set of vectors forms a basis for n-space, any $n \times 1$ vector $\mathbf{y}$ will be expressible as an exact linear combination of the columns of that $\mathbf{X}$. But of course "fitting the data" is not a proper objective of research using the CR model.

• The fact that R^2 tends to increase as additional explanatory variables are included leads some researchers to report an *adjusted* (or "corrected") *coefficient of determination*, which discounts the fit when k is large relative to n. This measure, referred to as $\bar{R}^2$ (read as "R bar squared"), is defined via

$$1 - \bar{R}^2 = (n - 1)(1 - R^2)/(n - k),$$

which inflates the unexplained proportion and hence deflates the explained proportion. There is no strong argument for using this particular adjustment: for example, $(1 - k/n)R^2$ would have a similar effect. It may well be preferable to report R^2, n, and k, and let readers decide how to allow for n and k.

• The adjusted coefficient of determination may be written explicitly as

$$\bar{R}^2 = 1 - \left[\sum_i e_i^2/(n - k)\right] \Big/ \left[\sum_i (y_i - \bar{y})^2/(n - 1)\right]. \tag{16.9}$$

FIFTH EDITION

ECONOMETRIC ANALYSIS

William H. Greene

New York University

Upper Saddle River, New Jersey 07458

is nonlinear. The reason is that if we write it in the form of (7-12), we fail to account for the condition that β_4 equals $\beta_2\beta_3$, which is a **nonlinear restriction.** In this model, the three parameters α, β, and γ are **overidentified** in terms of the four parameters β_1, β_2, β_3, and β_4. Unrestricted least squares estimates of β_2, β_3, and β_4 can be used to obtain two estimates of each of the underlying parameters, and there is no assurance that these will be the same.

7.4 MODELING AND TESTING FOR A STRUCTURAL BREAK

One of the more common applications of the F test is in tests of **structural change.**[8] In specifying a regression model, we assume that its assumptions apply to all the observations in our sample. It is straightforward, however, to test the hypothesis that some of or all the regression coefficients are different in different subsets of the data. To analyze a number of examples, we will revisit the data on the U.S. gasoline market[9] that we examined in Example 2.3. As Figure 7.5 following suggests, this market behaved in predictable, unremarkable fashion prior to the oil shock of 1973 and was quite volatile thereafter. The large jumps in price in 1973 and 1980 are clearly visible, as is the much greater variability in consumption. It seems unlikely that the same regression model would apply to both periods.

7.4.1 DIFFERENT PARAMETER VECTORS

The gasoline consumption data span two very different periods. Up to 1973, fuel was plentiful and world prices for gasoline had been stable or falling for at least two decades. The embargo of 1973 marked a transition in this market (at least for a decade or so), marked by shortages, rising prices, and intermittent turmoil. It is possible that the entire relationship described by our regression model changed in 1974. To test this as a hypothesis, we could proceed as follows: Denote the first 14 years of the data in $\mathbf{y}$ and $\mathbf{X}$ as $\mathbf{y}_1$ and $\mathbf{X}_1$ and the remaining years as $\mathbf{y}_2$ and $\mathbf{X}_2$. An unrestricted regression that allows the coefficients to be different in the two periods is

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}. \tag{7-13}$$

Denoting the data matrices as $\mathbf{y}$ and $\mathbf{X}$, we find that the unrestricted least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1'\mathbf{y}_1\\ \mathbf{X}_2'\mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix}, \tag{7-14}$$

which is least squares applied to the two equations separately. Therefore, the total sum of squared residuals from this regression will be the sum of the two residual sums of

[8] This test is often labeled a **Chow test,** in reference to Chow (1960).

[9] The data are listed in Appendix Table A6.1.

AN INTRODUCTION TO CLASSICAL ECONOMETRIC THEORY

Paul A. Ruud
University of California, Berkeley

New York • Oxford
OXFORD UNIVERSITY PRESS
2000

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{IV}} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \\ &= \boldsymbol{\beta}_0 + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta}_0 + \left(\mathrm{E}_{T|2}[\mathbf{z}_t\mathbf{x}_t']\right)^{-1}\mathrm{E}_{T|2}[\mathbf{z}_t\varepsilon_t]\end{aligned} \tag{20.20}$$

The z_{tk} $(k = 1, \ldots, K)$ are so-called instrumental variables (or *instruments*). By inspection we see that this IV estimator is consistent if the instrumental variables z_{tk} exhibit two characteristics: (1) $(\mathrm{E}_{T|2}[\mathbf{z}_t\mathbf{x}_t'])^{-1}$ converges in probability and (2) the z_{tk} are orthogonal to ε_t so that $\mathrm{E}_{T|2}[\mathbf{z}_t\varepsilon_t]$ converges in probability to a vector of zeros. In this chapter we describe how such instrumental variables arise in several models with latent variables.

20.2 LATENT VARIABLE MODELS

Econometrics is filled with latent variable models such as the one we have just studied. In this section we introduce several other important examples. All lead to the linear specification

$$y_n = \mathbf{x}_n'\boldsymbol{\beta}_0 + \varepsilon_n \qquad (n = 1, \ldots, N) \tag{20.21}$$

where ε_n is an unobserved, or latent, random variable. It is not merely the residual $y_n - \mathrm{E}[y_n \mid \mathbf{x}_n]$ defined by the choice of conditioning variables in $\mathbf{x}_n$.[8] In each model that we describe, at least one of the explanatory variables in $\mathbf{x}_n$ is correlated with ε_n so that $\mathrm{E}[\varepsilon_n \mid \mathbf{x}_n]$ is a function of $\mathbf{x}_n$ and, therefore, not zero. This in turn implies that $\mathrm{E}[y_n \mid \mathbf{x}_n] \neq \mathbf{x}_n'\boldsymbol{\beta}_0$ and that the OLS fit of y_n to $\mathbf{x}_n$ will yield inconsistent estimators of $\boldsymbol{\beta}_0$.

Researchers often call ε_n a *disturbance* or *error term*. This seems appropriate when assumptions are made directly about the behavior of the latent ε_n, rather than about the observable variables y_n and $\mathbf{x}_n$ only. The next example, measurement errors in the explanatory variables, contains such assumptions and describes simply a fundamental problem in actual empirical work.

EXAMPLE 20.1 (Errors in Variables)

Suppose that we are interested in the regression function

$$\mathrm{E}[y_n \mid \mathbf{x}_n^*] = \mathbf{x}_n^{*\prime}\boldsymbol{\beta}_0$$

but some of the explanatory variables in $\mathbf{x}_n^*$ are not observable. Such economic variables as expected price inflation, transaction costs, ability or productivity of an employee, and supply or demand shocks are examples of such latent variables. But we may observe *proxy variables*: actual inflation might take the place of price expectations or an individual's IQ might serve as an imperfect measure of cognitive ability. We denote these proxy variables with $\mathbf{x}_n$ and let

$$\mathbf{x}_n = \mathbf{x}_n^* + \boldsymbol{\upsilon}_n$$

where $\boldsymbol{\upsilon}_n$ denotes the measurement errors in the proxy variables. It is simplest to suppose that $\mathrm{E}[\boldsymbol{\upsilon}_n] = \mathbf{0}$ so that the proxy variables exhibit no systematic bias. We also assume that the $\boldsymbol{\upsilon}_n$ are

[8] The term "latent" has a narrower meaning for some writers. They require that latent variables are not implicit functions of observable variables. Within this definition, $\varepsilon_n = y_n - \mathbf{x}_n'\boldsymbol{\beta}_0$ is not latent. Instead, ε_n would be called "unmeasured." See Aigner et al. (1984, p. 1323) and the reference they cite, Bentler (1982).

uncorrelated with both the x^*_{nk} $(k = 1, \ldots, K)$ and $u_n \equiv y_n - \mathbf{x}^{*\prime}_n \boldsymbol{\beta}_0$. Then the feasible regression relationship is given by

$$\begin{aligned} y_n &= (\mathbf{x}_n - \boldsymbol{\upsilon}_n)' \boldsymbol{\beta}_0 + u_n \\ &= \mathbf{x}'_n \boldsymbol{\beta}_0 + \varepsilon_n \end{aligned}$$

where the latent disturbance $\varepsilon_n \equiv u_n - \boldsymbol{\upsilon}'_n \boldsymbol{\beta}_0$ is correlated with $\mathbf{x}_n$ because $\boldsymbol{\upsilon}_n$ is a latent component of $\mathbf{x}_n$.

Another explanation for correlation between the explanatory variables and the residual term is a *system of simultaneous equations*. Such models are common in econometrics, in part because multivariate optimization and equilibrium are prevalent features of economic models.

EXAMPLE 20.2 (Simultaneous Equations)

Consider the simple market model in which there is a supply function

$$q_{sn} = \mathbf{x}'_{s1n} \boldsymbol{\beta}_{0s1} + \beta_{0s2} p_n + \varepsilon_{sn} \tag{20.22}$$

for the aggregate supply of a good q_{sn} available at market price p_n and a demand function

$$q_{dn} = \mathbf{x}'_{d1n} \boldsymbol{\beta}_{0d1} + \beta_{0d2} p_n + \varepsilon_{dn} \tag{20.23}$$

for the aggregate demand q_{dn} at market price p_n. The ε_{sn} and ε_{dn} are latent random disturbance terms. We partition the observable explanatory variables $\mathbf{x}_{sn} \equiv \left[\mathbf{x}'_{s1n},\ p_n\right]'$ and $\mathbf{x}_{dn} \equiv \left[\mathbf{x}'_{d1n},\ p_n\right]'$ to distinguish the market price from $\mathbf{x}_{s1n}$ and $\mathbf{x}_{d1n}$. These are assumed to be predetermined, capturing exogenous shifts in the supply and demand functions. Let $\mathbf{x}_n \equiv \left[\mathbf{x}'_{s1n},\ \mathbf{x}'_{d1n}\right]'$ and $(\varepsilon_{sn},\ \varepsilon_{dn})$ be i.i.d. random variables with finite fourth moments and $\mathrm{E}[\varepsilon_{sn} \mid \mathbf{x}_n] = \mathrm{E}[\varepsilon_{dn} \mid \mathbf{x}_n] = 0$.

In equilibrium, the market price will clear the market so that the observed quantity transacted, y_n, equals both the desired supply and the desired demand:

$$y_n = q_{sn} = q_{dn}$$

Therefore,

$$y_n = \mathbf{x}'_{s1n} \boldsymbol{\beta}_{0s1} + \beta_{0s2} p_n + \varepsilon_{sn} = \mathbf{x}'_{d1n} \boldsymbol{\beta}_{0d1} + \beta_{0d2} p_n + \varepsilon_{dn}$$

and we can solve this system of simultaneous equations for the equilibrium price

$$p_n = \frac{1}{\beta_{0s2} - \beta_{0d2}} \left(\mathbf{x}'_{d1n} \boldsymbol{\beta}_{0d1} - \mathbf{x}'_{s1n} \boldsymbol{\beta}_{0s1} + \varepsilon_{dn} - \varepsilon_{sn}\right) \tag{20.24}$$

given that $\beta_{0d2} < 0 < \beta_{0s2}$.[9] It follows that the explanatory variable p_n in both demand and supply equations will be correlated with both ε_{sn} and ε_{dn}. Because p_n and y_n are jointly determined by the interaction of supply and demand, p_n is a function of the latent disturbance terms in both supply and demand functions. OLS estimation of either (20.22) or (20.23) will yield biased and inconsistent estimates.

Our final example is the most direct. For discussions of estimation, we will also use this example to describe all of the previous examples as well.

[9] We give a detailed description of simultaneous equations models in Chapter 26.

$\tau_0 \equiv [\tau_{0k}]'$, Π_0 is the $K_1 \times (K - K_1)$ matrix of coefficients $[\pi_{0k};\ k = K_1 + 1, \ldots, K]$, and we partition the K elements of $\mathbf{x}_n$ into K_1 and $K - K_1$ elements $[\mathbf{x}_{1n}', \mathbf{x}_{2n}']'$, respectively. The first term in the coefficient vector of $\mathbf{x}_{1n}$ in (20.26) is γ_{01}, the coefficient vector of $\mathbf{x}_{1n}$ in $\mathrm{E}^*[y_n \mid \mathbf{x}_n]$. The second term is the product of the coefficients of $\mathbf{x}_{1n}$ in $\mathrm{E}^*[\mathbf{x}_{2n} \mid \mathbf{x}_{1n}]$ and the coefficient vector of $\mathbf{x}_{2n}$ in $\mathrm{E}^*[y_n \mid \mathbf{x}_n]$. This term is an adjustment to γ_{01} that takes into account predictable differences in y_n that are associated with $\mathbf{x}_{2n}$ and that $\mathbf{x}_{1n}$ can also capture through its power to predict $\mathbf{x}_{2n}$.

These two components are analogous to the components of the total derivative of a function of two variables $f(x_1, x_2)$ with respect to the first variable:

$$\frac{df(x_1, x_2)}{dx_1} = \frac{\partial f(x_1, x_2)}{\partial x_1} + \frac{dx_2}{dx_1}\frac{\partial f(x_1, x_2)}{\partial x_2}$$

The first term is the *ceteris paribus* change in f for a change in x_1 and the second term is the product of the *ceteris paribus* change in f for a change in x_2 and the change in x_2 accompanying a change in x_1.[12] In this analogy, we interpret the function f as $\mathrm{E}^*[y \mid \mathbf{x}_{1n}, \mathbf{x}_{2n}]$. The derivative dx_2/dx_1 corresponds to Π_0.

Suppose now that $\mathrm{E}[y_n \mid \mathbf{x}_n] = \mathbf{x}_{1n}'\boldsymbol{\beta}_{01} + \beta_{02}x_{2n}$ but that we do not include one variable, x_{2n}, in the OLS estimation of $\boldsymbol{\beta}_{01}$. Lemmas 20.1 and 20.2 indicate that when we regress y_n on $\mathbf{x}_{1n}$ alone, the OLS fitted coefficients $\hat{\boldsymbol{\beta}}_{\mathrm{R}1}$ will generally converge in probability to

$$\gamma_{01} = \boldsymbol{\beta}_{01} + \Pi_0\beta_{02} \tag{20.27}$$

Therefore, we can interpret the probability limit of the elements of $\hat{\boldsymbol{\beta}}_{1\mathrm{R}}$ as the sum of two terms: the direct change in the expected value of y_n associated with a change in $\mathbf{x}_{1nk}$, β_{01k}, plus an indirect change in the expected value of y_n associated with changes in x_{2n}, $\pi_{0k}\beta_{02}$, for each k.

If there were no correlation between x_{2n} and $\mathbf{x}_{1n}$, then the latter would have no (linear) predictive power for x_{2n} and there would be no indirect effect because $\Pi_0 = \mathbf{0}$. We would estimate only $\boldsymbol{\beta}_{01}$. This also occurs, of course, if $\beta_{02} = 0$. Otherwise, to the extent that linear prediction allows, the OLS procedure fits the variation in y_n with $\mathbf{x}_{1n}$ as well as possible, leading to the addition of the indirect effects to the direct ones in the probability limit of $\hat{\boldsymbol{\beta}}_{\mathrm{R}1}$.

Note that in general *all* of the estimated coefficients may be affected by the omission of an explanatory variable. The bias and inconsistency are not limited only to the coefficients of those explanatory variables that are correlated with the omitted variable. One can see this algebraically in $\Pi_0 = (\mathrm{Var}[\mathbf{x}_{1n}])^{-1}\,\mathrm{Cov}[\mathbf{x}_{1n}, x_{2n}]$. The covariance term is premultiplied by the inverse of a variance matrix, which potentially spreads any nonzero covariance across all elements of the matrix product. This phenomenon represents the effects of MSE optimization: as one coefficient adjusts to account for a missing explanatory variable, the other coefficients adjust in turn to account for this. As a result, the effects of an omitted explanatory variable generally vitiate estimates of every coefficient.

In special cases, it is possible to predict the effects of the omitted explanatory variable.

EXAMPLE 20.4 (Errors in Variables)

The model of errors in explanatory variables predicts a definite direction for inconsistency in simple regression. Researchers commonly use this prediction to interpret their estimates of multivariate regressions. Specializing the model described in Example 20.1 to simple regression, we have

[12] Recall Exercise 3.8.

$$y_n = \beta_0 x_n - \beta_0 \upsilon_n + u_n$$

so that an OLS fit of y_n to x_n implicitly omits υ_n. Because

$$\mathrm{Cov}[u_n, \upsilon_n] = \mathrm{Cov}[x_n^*, u_n] = \mathrm{Cov}[x_n^*, \upsilon_n] = 0 \tag{20.28}$$

it follows that

$$\mathrm{E}[x_n \upsilon_n] = \mathrm{Cov}[x_n, \upsilon_n] = \mathrm{Var}[\upsilon_n] > 0 \tag{20.29}$$

In words, the observable proxy variable x_n and its measurement error υ_n are positively correlated. Therefore

$$\pi_0 = \frac{\mathrm{Cov}[x_n, \upsilon_n]}{\mathrm{E}[x_n^2]} = \frac{\mathrm{Var}[\upsilon_n]}{\mathrm{E}[x_n^2]} > 0$$

in $\mathrm{E}^*[\upsilon_n \mid x_n] = x_n \pi_0$ and the inconsistency in $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, which is $-\pi_0 \beta_0$, will have the opposite sign of β_0.

We can also show that the inconsistency is not so large that the *sign* of plim $\hat{\beta}_{\mathrm{OLS}}$ differs from that of β_0: using (20.28),

$$\mathrm{E}[x_n^2] = \mathrm{E}[x_n^{*2}] + \mathrm{Var}[\upsilon_n] \quad \Rightarrow \quad 0 < \pi_0 < 1 \tag{20.30}$$

Therefore, errors in an explanatory variables shrink the probability limit toward zero relative to the coefficient:

$$\text{plim } \hat{\beta}_{\mathrm{OLS}} = \beta_0 (1 - \pi_0) \tag{20.31}$$

In other words, it diminishes the apparent influence of a latent explanatory variable. This is exactly what common sense suggests measurement error should do.

Figure 20.1 gives a graphic description of this example for the case in which $\mathrm{E}[x_n^*] = 0$. The variance ellipsoid for (x_n^*, y_n) is labeled $\mathbb{V}^*$. It is framed by a dashed box two standard deviations

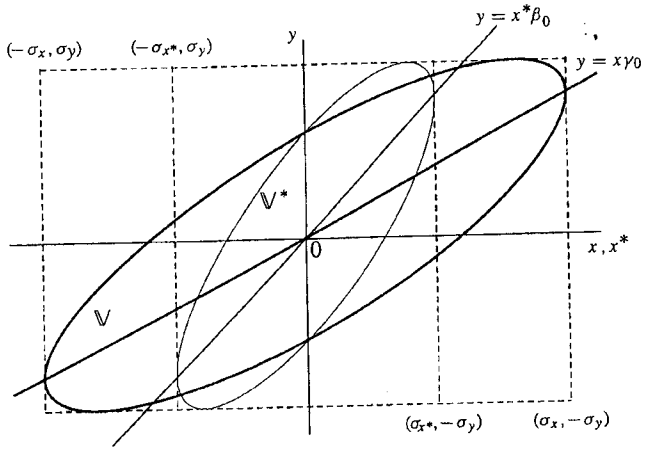

Figure 20.1 Errors in variables.

on each side, as in Figures 7.3–7.5. The MMSE linear prediction line $y = x^*\beta_0$ for this ellipsoid appears as a solid line. As in Figure 7.8, this line intersects the vertical tangents to the variance ellipsoid. The variance ellipsoid for (x_n, y_n) is thicker and is labeled $\mathbb{V}$. It is framed by a box that is the same height as that for $\mathbb{V}^*$ because the standard deviation of y_n is constant. The box framing $\mathbb{V}$ is wider than that for $\mathbb{V}^*$ because the variance of x_n is larger than the variance of x_n^* by $\text{Var}[\upsilon_n]$, as in (20.30). As a result, the line of vertical tangent points to $\mathbb{V}$ must have a smaller slope, yet the slope will not change sign; and that thick line is the MMSE linear prediction line $y = x\gamma_0$.

We cannot offer such simple descriptions of the inconsistency of OLS in the dynamic regression or simultaneous equations examples. Instead, we characterize the explanatory variables that have been omitted by finding MMSE linear predictors for each case. In all of our examples, a latent variable model describes the cause of $\text{E}[\varepsilon_n \mid \mathbf{x}_n] \neq 0$ despite the general structure in which $y_n = \mathbf{x}_n'\boldsymbol{\beta}_0 + \varepsilon_n$ and $\text{E}[\varepsilon_n] = 0$. In each case, interest focuses on $\mathbf{x}_n'\boldsymbol{\beta}_0$ but this is not the conditional mean of y_n given $\mathbf{x}_n$. We will reformulate every cause as an inability to condition the mean of y_n on all of the necessary explanatory variables. A critical explanatory variable is latent and, for this reason, omitted.

For errors in explanatory variables (Example 20.1), this point is trivial. If one could include in the conditioning set the measurement error υ_n, then

$$\text{E}[y_n \mid \mathbf{x}_n, \upsilon_n] = \mathbf{x}_n'\boldsymbol{\beta}_0 - \upsilon_n'\boldsymbol{\beta}_0$$

would be specified well enough to estimate $\boldsymbol{\beta}_0$ with OLS. But that is simply stating that if $\mathbf{x}_n^*$ were observable we could regress y_n on $\mathbf{x}_n^*$. For the dynamic regression in the previous section, this point is not trivial.

EXAMPLE 20.5 (Dynamic Regression)

We saw in (20.15) and (20.18) that the success of GLS estimation implicitly rests on the inclusion of additional variables in the regression equation. That is, if we expand the conditioning set to include $\mathbf{x}_{t-1} \equiv [\mathbf{x}_{1,t-1}', y_{t-2}]'$ then we obtain

$$\text{E}[y_t \mid \mathbf{x}_t, \mathbf{x}_{t-1}] = \mathbf{x}_t'\boldsymbol{\beta}_0 + \phi_0(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta}_0) \tag{20.32}$$

Alternatively, this conditional mean corrects $\mathbf{x}_t'\boldsymbol{\beta}_0$ for the missing latent explanatory variable $\varepsilon_{t-1} = y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta}_0$:

$$\text{E}[y_t \mid \mathbf{x}_t, \mathbf{x}_{t-1}] = \mathbf{x}_t'\boldsymbol{\beta}_0 + \phi_0\varepsilon_{t-1} = \text{E}[y_t \mid \mathbf{x}_t, \varepsilon_{t-1}] \tag{20.33}$$

This conditional mean also suggests a consistent estimator of $\boldsymbol{\beta}_0$. If we expand (20.32), then

$$\begin{aligned}\text{E}[y_t \mid \mathbf{x}_t, \mathbf{x}_{t-1}] = {} & \mathbf{x}_{1t}'\boldsymbol{\beta}_{01} + (\beta_{02} + \phi_0)\, y_{t-1} \\ & + \mathbf{x}_{1,t-1}'(-\phi_0 \cdot \boldsymbol{\beta}_{01}) + (\phi_0\beta_{02})\, y_{t-2}\end{aligned} \tag{20.34}$$

can be estimated with OLS. The fitted coefficients of $\mathbf{x}_{1t}$ are consistent estimators of $\boldsymbol{\beta}_{01}$ and the coefficients of $\mathbf{x}_{1,t-1}$ are consistent estimators of $\phi_0 \cdot \boldsymbol{\beta}_{01}$. Hence, ϕ_0 is consistently estimated by the ratios of these coefficients. This in turn implies that we may estimate β_{02} with the fitted coefficient of y_{t-1} minus the estimator of ϕ_0 or the fitted coefficient of y_{t-2} divided by the estimator of ϕ_0.[13]

[13] This ratio will not be a reliable estimator if $\phi_0 = 0$. For this reason, the difference estimator is preferred. The same issue arises in the estimation of ϕ_0. Consistent estimation requires that the element of $\boldsymbol{\beta}_{01}$ be nonzero.

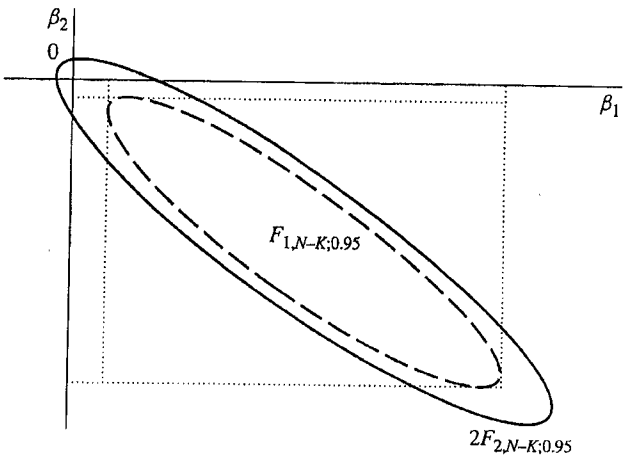

Figure 11.4 Joint versus marginal statistical significance.

β_2 whereas the p value of the F statistic is 0.9317. At the 5% level of significance, the t statistics are statistically significant and the F statistic is not, but practically speaking this distinction is too sharp.

Second, the sound bite "significant estimated coefficient" is often heard as "fairly large effect," rather than "the interval estimator does not contain zero." The phrase "insignificant estimated coefficient" may sound like "negligible effect." But the fact that the interval estimator contains zero says nothing about whether the interval estimator contains large coefficient values as well. We might ascribe such failures in communication to confusion of hypothesis testing with interval estimation.[10] The *statistical* significance of t statistics concerns particular values of the coefficients. One evaluates the *qualitative* significance of the estimated coefficients using interval estimation, evaluating the entire range of likely values.

Put another way, an "insignificant" t statistic can occur for two reasons: the estimated coefficient may be qualitatively close to zero *or* the estimated standard error may be very large. In the latter case, a statistically imprecise estimate supports both a small and a large effect in the population; it is uninformative.

11.3.3 Optimal Power and the F Test

The F test does not possess an optimal power function. One might expect a result analogous to earlier results on the efficiency of OLS estimators, especially in light of those results. Nevertheless, just as two estimators may not be ordered by relative efficiency, two hypothesis tests may have power functions that cross in the parameter space of the alternative hypothesis. Hence, the very existence of a most powerful test would be an interesting result and such tests do occur in special cases. However, testing $\mathbf{R}\boldsymbol{\beta}_0 = \mathbf{r}$ in the normal regression model is not one of those cases. The

[10] We also recognize that there may be other explanations for this usage.

(except p_n) and $\mathbf{z}_n = \left[\mathbf{x}'_{s1n}, x_{dnk}\right]'$ such that $\mathrm{E}[\mathbf{z}_n\mathbf{x}'_n]$ is nonsingular to construct an IV estimator for $\boldsymbol{\beta}_{0s} = \left[\boldsymbol{\beta}'_{0s1}, \beta_{0s2}\right]'$.

In both of these examples, there is generally an infinite number of IV estimators. We can also consider general functions of the valid instrumental variables. For example, the family $\mathbf{Z} = \left\{\left[\mathbf{x}'_{s1n}, f(\mathbf{x}_{s1n}, \mathbf{x}_{d1n})\right]'\right\}'$ for various functions f contains potential instrument matrices for the supply equation of the simultaneous market system. The necessary orthogonality will still hold, so that we are constrained only by the requirement that $\mathbf{Z}'\mathbf{X}$ be nonsingular. This is critical for the errors in explanatory variables example.

EXAMPLE 20.9 (Errors in Variables)

Example 20.1 with errors in the explanatory variables comes with no "extra" variables comparable to $\mathbf{x}_{1,t-1}$ or $\mathbf{x}_{d1n}$ in the previous two examples. However, nonlinear functions of the observed explanatory variables may still provide a valid instrument matrix $\mathbf{Z}$ under well-specified circumstances. To illustrate, consider a case with three explanatory variables, one of which is measured with error. Let

$$\mathrm{E}[y_n \mid x_{2n}, x^*_{n3}] = \beta_{01} + \beta_{02}x_{2n} + \beta_{03}x^*_{n3}$$

and $x_{n3} = x^*_{n3} + \upsilon_n$ be the variable measured with error. Suppose that both x_{2n} and x^2_{2n} are correlated with x^*_{n3}. Because x^2_{2n} is uncorrelated with υ_n, $\mathbf{z}_n = \left[1, x_{2n}, x^2_{2n}\right]$ would generally be a valid list of instrumental variables.

However, most researchers would not accept such an estimator for empirical use. The reason is that there is another, plausible interpretation of the estimated coefficients. Because we do not know that the conditional mean is a linear function of x_{2n} in actual applications, we might consider the possibility that x^2_{2n} should also be included as an RHS explanatory variable. But if it is, then we will need a third instrumental variable and we are back to looking for another function of x_{2n} to serve as an instrumental variable. Because such an argument can be made for any function of x_{2n}, the use of nonlinear functions as instrumental variables for the errors-in-variables problem is widely viewed with suspicion.

Thus, we must recognize that not all problems have IV solutions. There are situations in which the parameters of the model cannot be estimated. In this characteristic, these situations are similar to exact multicollinearity among the explanatory variables. When there is exact multicollinearity, the matrix $\mathbf{X}'\mathbf{X}$ is singular. When there is no consistent IV estimator, one cannot construct a nonsingular $\mathbf{Z}'\mathbf{X}$ from the available information. In both cases, the parameters of the model are not identified.

20.5 TWO-STAGE LEAST SQUARES

The IV estimators that latent models suggest are often more specific than a list of possible instrumental variables. The models may also offer insight into the particular functions of the available variables that provide appealing estimators. In this section, we delve more deeply into the the example of simultaneous equations. In such linear systems, an intuitively attractive instrumental variable for p_n is $\mathbf{x}'_n\hat{\boldsymbol{\pi}}_p$, the OLS fitted value from the regression of p_n on *all* of the

Introduction to Econometrics

James H. Stock
HARVARD UNIVERSITY

Mark W. Watson
PRINCETON UNIVERSITY

Boston San Francisco New York
London Toronto Sydney Tokyo Madrid
Mexico City Munich Paris Cape Town Hong Kong Montreal

One source of discrete breaks in macroeconomic data is a major change in macroeconomic policy. For example, the breakdown of the Bretton Woods system of fixed exchange rates in 1972 produced the break in the time series behavior of the \$/£ exchange rate that is evident in Figure 12.2b. Prior to 1972, the exchange rate was essentially constant, with the exception of a single devaluation in 1968 in which the official value of the pound, relative to the dollar, was decreased. In contrast, since 1972 the exchange rate has fluctuated over a very wide range.

Breaks also can occur more slowly as the population regression evolves over time. For example, such changes can arise because of slow evolution of economic policy and ongoing changes in the structure of the economy. The methods for detecting breaks described in this section can detect both types of breaks, distinct changes and slow evolution.

Problems caused by breaks. If a break occurs in the population regression function during the sample, then the OLS regression estimates over the full sample will estimate a relationship that holds "on average," in the sense that the estimate combines the two different periods. Depending on the location and the size of the break, the "average" regression function can be quite different than the true regression function at the end of the sample, and this leads to poor forecasts.

Testing for Breaks

One way to detect breaks is to test for discrete changes, or breaks, in the regression coefficients. How this is done depends on whether the date of the suspected break (the **break date**) is known or not.

Testing for a break at a known date. In some applications you might suspect that there is a break at a known date. For example, if you are studying international trade relationships using data from the 1970s, you might hypothesize that there is a break in the population regression function of interest in 1972 when the Bretton Woods system of fixed exchange rates was abandoned in favor of floating exchange rates.

If the date of the hypothesized break in the coefficients is known, then the null hypothesis of no break can be tested using a binary variable interaction regression of the type discussed in Chapter 6 (Key Concept 6.4). To keep things simple, consider an ADL(1,1) model, so there is an intercept, a single lag of Y_t, and a single lag of X_t. Let τ denote the hypothesized break date and let $D_t(\tau)$ be a binary variable that equals zero before the break date and one after, so $D_t(\tau) = 0$

if $t \leq \tau$ and $D_t(\tau) = 1$ if $t > \tau$. Then the regression including the binary break indicator and all interaction terms is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_0 D_t(\tau) + \gamma_1 [D_t(\tau) \times Y_{t-1}] + \gamma_2 [D_t(\tau) \times X_{t-1}] + u_t. \quad (12.35)$$

If there is not a break, then the population regression function is the same over both parts of the sample so the terms involving the break binary variable $D_t(\tau)$ do not enter Equation (12.35). That is, under the null hypothesis of no break, $\gamma_0 = \gamma_1 = \gamma_2 = 0$. Under the alternative hypothesis that there is a break, then the population regression function is different before and after the break date τ, in which case at least one of the γ's is nonzero. Thus the hypothesis of a break can be tested using the F-statistic that tests the hypothesis that $\gamma_0 = \gamma_1 = \gamma_2 = 0$ against the hypothesis that at least one of the γ's is nonzero. This is often called a Chow test for a break at a known break date, named for its inventor, Gregory Chow (1960).

If there are multiple predictors or more lags, then this test can be extended by constructing binary variable interaction variables for all the regressors and testing the hypothesis that all the coefficients on terms involving $D_t(\tau)$ are zero.

This approach can be modified to check for a break in a subset of the coefficients by including only the binary variable interactions for that subset of regressors of interest.

Testing for a break at an unknown break date Often the date of a possible break is unknown or known only within a range. Suppose, for example, you suspect that a break occurred sometime between two dates, τ_0 and τ_1. The Chow test can be modified to handle this by testing for breaks at all possible dates τ in between τ_0 and τ_1, then using the largest of the resulting F-statistics to test for a break at an unknown date. This modified Chow test is variously called the **Quandt likelihood ratio (QLR) statistic** (Quandt, 1960) (the term we shall use) or, more obscurely, the **sup-Wald statistic**.

Because the QLR statistic is the largest of many F-statistics, its distribution is not the same as an individual F-statistic. Instead, the critical values for the QLR statistic must be obtained from a special distribution. Like the F-statistic, this distribution depends on the number of restrictions being tested, q, that is, the number of coefficients (including the intercept) that are being allowed to break, or change, under the alternative hypothesis. The distribution of the QLR statistic also depends on τ_0/T and τ_1/T, that is, on the endpoints, τ_0 and τ_1, of the subsample over which the F-statistics are computed, expressed as a fraction of the total sample size.

Key Concept 7.2

Which Variables to Add to My Regression?

If you include another variable in your multiple regression, you will eliminate the possibility of omitted variable bias from excluding that variable but the variance of the estimator of the coefficients of interest can increase. Here are some guidelines to help you decide whether to include an additional variable:

1. Be specific about the coefficient or coefficients of interest.
2. Use *a priori* reasoning to identify the most important potential sources of omitted variable bias, leading to a base specification and some "questionable" variables.
3. Test whether additional questionable variables have nonzero coefficients.
4. Provide "full disclosure" representative tabulations of your results so that others can see the effect of including the questionable variables on the coefficient(s) of interest. Do your results change if you include a questionable variable?

Errors-in-Variables

Suppose that in our regression of test scores against the student-teacher ratio we had inadvertently mixed up our data, so that we ended up regressing test scores for fifth graders on the student-teacher ratio for tenth graders in that district. Although the student-teacher ratio for elementary school students and tenth graders might be correlated, they are not the same, so this mix-up would lead to bias in the estimated coefficient. This is an example of **errors-in-variables bias** because its source is an error in the measurement of the independent variable. This bias persists even in very large samples, so that the OLS estimator is inconsistent if there is measurement error.

There are many possible sources of measurement error. If the data are collected through a survey, a respondent might give the wrong answer. For example, one question in the Current Population Survey involves last year's earnings. A respondent might not know his exact earnings, or he might misstate it for some other reason. If instead the data are obtained from computerized administrative records, there might have been typographical errors when the data were first entered.

To see that errors-in-variables bias results in correlation between the regressor and the error term, suppose there is a single regressor X_i (say, actual income)

but that X_i is measured imprecisely by $\tilde{X}_i$ (the respondent's estimate of income). Because $\tilde{X}_i$, not X_i, is observed, the regression equation actually estimated is the one based on $\tilde{X}_i$. Written in terms of the imprecisely measured variable $\tilde{X}_i$, the population regression equation $Y_i = \beta_0 + \beta_1 X_i + u_i$ is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1(X_i - \tilde{X}_i) + u_i] \\ &= \beta_0 + \beta_1 \tilde{X}_i + v_i, \end{aligned} \tag{7.1}$$

where $v_i = \beta_1(X_i - \tilde{X}_i) + u_i$. Thus, the population regression equation written in terms of $\tilde{X}_i$ has an error term that contains the difference between X_i and $\tilde{X}_i$. If this difference is correlated with the measured value $\tilde{X}_i$, then the regressor $\tilde{X}_i$ will be correlated with the error term and $\hat{\beta}_1$ will be biased and inconsistent.

The precise size and direction of the bias in $\hat{\beta}_1$ depends on the correlation between $\tilde{X}_i$ and $(X_i - \tilde{X}_i)$. This correlation depends, in turn, on the specific nature of the measurement error.

As an example, suppose that the survey respondent provides their best guess or recollection of the actual value of the independent variable X_i. A convenient way to represent this mathematically is to suppose that the measured value of X_i equals the actual, unmeasured value, plus a purely random component, w_i. Accordingly, the measured value of the variable, denoted by $\tilde{X}_i$, is $\tilde{X}_i = X_i + w_i$. Because the error is purely random, we might suppose that w_i has mean zero and variance σ_w^2 and is uncorrelated with X_i and the regression error u_i. Under this assumption, a bit of algebra[2] shows that $\hat{\beta}_1$ has the probability limit

$$\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\beta_1. \tag{7.2}$$

That is, if the measurement imprecision has the effect of simply adding a random element to the actual value of the independent variable, then $\hat{\beta}_1$ is inconsistent. Because the ratio $\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$ is less than one, $\hat{\beta}_1$ will be biased towards zero, even in large samples. In the extreme case that the measurement error is so large that essentially no information about X_i remains, the ratio of the variances in the final expression in Equation (7.2) is zero and $\hat{\beta}_1$ converges in probability to zero. In the other extreme, when there is no measurement error, $\sigma_w^2 = 0$ so $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

[2]Under this measurement error assumption, $v_i = \beta_1(X_i - \tilde{X}_i) + u_i = -\beta_1 w_i + u_i$, $\text{cov}(\tilde{X}_i, u_i) = 0$, and $\text{cov}(\tilde{X}_i, w_i) = \text{cov}(X_i + w_i, w_i) = \sigma_w^2$, so $\text{cov}(\tilde{X}_i, v_i) = -\beta_1 \text{cov}(\tilde{X}_i, w_i) + \text{cov}(\tilde{X}_i, u_i) = -\beta_1 \sigma_w^2$. Thus, from Equation (5.1), $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \sigma_w^2 / \sigma_{\tilde{X}}^2$. Now $\sigma_{\tilde{X}}^2 = \sigma_X^2 + \sigma_w^2$, so $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \sigma_w^2 / (\sigma_X^2 + \sigma_w^2) = [\sigma_X^2 / (\sigma_X^2 + \sigma_w^2)]\beta_1$.

Although the result in Equation (7.2) is specific to this particular type of measurement error, it illustrates the more general proposition that if the independent variable is measured imprecisely then the OLS estimator is biased, even in large samples. Errors-in-variables bias is summarized in Key Concept 7.3.

Solutions to errors-in-variables bias. The best way to solve the errors-in-variables problem is to get an accurate measure of X. If this is impossible, however, there are econometric methods that can be used to mitigate errors-in-variables bias.

One such method is instrumental variables regression. This relies on having another variable (the "instrumental" variable) that is correlated with the actual value X_i but is uncorrelated with the measurement error. This method is studied in Chapter 10.

A second method is to develop a mathematical model of the measurement error and, if possible, to use the resulting formulas to adjust the estimates. For example, if a researcher believes that the measured variable is in fact the sum of the actual value and a random measurement error term, and if she knows or can estimate the ratio σ_w^2/σ_X^2, then she can use Equation (7.2) to compute an estimator of β_1 that corrects for the downward bias. Because this approach requires specialized knowledge about the nature of the measurement error, the details typically are specific to a given data set and its measurement problems and we shall not pursue this approach further in this textbook.

Sample Selection

Sample selection bias occurs when the availability of the data is influenced by a selection process that is related to the value of the dependent variable. This selection process can introduce correlation between the error term and the regressor, which leads to bias in the OLS estimator.

Sample selection that is unrelated to the value of the dependent variable does not introduce bias. For example, if data are collected from a population by simple random sampling, the sampling method (being drawn at random from the population) has nothing to do with the value of the dependent variable. Such sampling does not introduce bias.

Bias can be introduced when the method of sampling is related to the value of the dependent variable. An example of sample selection bias in polling was given in the box in Chapter 2. In that example, the sample selection method (randomly selected phone numbers of automobile owners) was related to the dependent variable (who the individual supported for president in 1936), because in 1936 car owners with phones were more likely to be Republicans.

THIRD EDITION

INTRODUCTORY ECONOMETRICS

A MODERN APPROACH

JEFFREY M. WOOLDRIDGE
Michigan State University

THOMSON
SOUTH-WESTERN

Australia · Canada · Mexico · Singapore · Spain · United Kingdom · United States

earn less than whites (for a given black population), and the opposite is true if the percentage of Hispanics is above 9.45%. Twelve of the 22 cities represented in the sample have Hispanic populations that are less than 6% of the total population. The largest percentage of Hispanics is about 31%.

How do we interpret these findings? We cannot simply claim discrimination exists against blacks and Hispanics, because the estimates imply that whites earn less than blacks and Hispanics in cities heavily populated by minorities. The importance of city composition on salaries might be due to player preferences: perhaps the best black players live disproportionately in cities with more blacks and the best Hispanic players tend to be in cities with more Hispanics. The estimates in (7.19) allow us to determine that some relationship is present, but we cannot distinguish between these two hypotheses.

Testing for Differences in Regression Functions across Groups

The previous examples illustrate that interacting dummy variables with other independent variables can be a powerful tool. Sometimes, we wish to test the null hypothesis that two populations or groups follow the same regression function, against the alternative that one or more of the slopes differ across the groups. We will also see examples of this in Chapter 13, when we discuss pooling different cross sections over time.

Suppose we want to test whether the same regression model describes college grade point averages for male and female college athletes. The equation is

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

where *sat* is SAT score, *hsperc* is high school rank percentile, and *tothrs* is total hours of college courses. We know that, to allow for an intercept difference, we can include a dummy variable for either males or females. If we want any of the slopes to depend on gender, we simply interact the appropriate variable with, say, *female*, and include it in the equation.

If we are interested in testing whether there is *any* difference between men and women, then we must allow a model where the intercept and all slopes can be different across the two groups:

$$cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u. \quad (7.20)$$

The parameter δ_0 is the difference in the intercept between women and men, δ_1 is the slope difference with respect to *sat* between women and men, and so on. The null hypothesis that *cumgpa* follows the same model for males and females is stated as

$$H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0. \quad (7.21)$$

If one of the δ_j is different from zero, then the model is different for men and women.

Using the spring semester data from the file GPA3.RAW, the full model is estimated as

$$\begin{aligned}\widehat{cumgpa} = 1.48 &- .353\,female + .0011\,sat + .00075\,female{\cdot}sat \\ &\quad (0.21)\ \ (.411)\ \ \ \ \ (.0002)\ \ \ (.00039) \\ &- .0085\,hsperc - .00055\,female{\cdot}hsperc + .0023\,tothrs \\ &\quad (.0014)\ \ \ \ \ (.00316)\ \ \ \ \ \ \ (.0009) \\ &- .00012\,female{\cdot}tothrs \\ &\quad (.00163) \\ n = 366,\ R^2 = .406,\ \bar{R}^2 &= .394. \end{aligned} \tag{7.22}$$

None of the four terms involving the female dummy variable is very statistically significant; only the *female·sat* interaction has a t statistic close to two. But we know better than to rely on the individual t statistics for testing a joint hypothesis such as (7.21). To compute the F statistic, we must estimate the restricted model, which results from dropping *female* and all of the interactions; this gives an R^2 (the restricted R^2) of about .352, so the F statistic is about 8.14; the p-value is zero to five decimal places, which causes us to soundly reject (7.21). Thus, men and women athletes do follow different GPA models, even though each term in (7.22) that allows women and men to be different is individually insignificant at the 5% level.

The large standard errors on *female* and the interaction terms make it difficult to tell exactly how men and women differ. We must be very careful in interpreting equation (7.22) because, in obtaining differences between women and men, the interaction terms must be taken into account. If we look only at the *female* variable, we would wrongly conclude that *cumgpa* is about .353 less for women than for men, holding other factors fixed. This is the estimated difference only when *sat*, *hsperc*, and *tothrs* are all set to zero, which is not close to being a possible scenario. At *sat* = 1,100, *hsperc* = 10, and *tothrs* = 50, the predicted *difference* between a woman and a man is $-.353 + .00075(1{,}100) - .00055(10) - .00012(50) \approx .461$. That is, the female athlete is predicted to have a GPA that is almost one-half a point higher than the comparable male athlete.

In a model with three variables, *sat*, *hsperc*, and *tothrs*, it is pretty simple to add all of the interactions to test for group differences. In some cases, many more explanatory variables are involved, and then it is convenient to have a different way to compute the statistic. It turns out that the sum of squared residuals form of the F statistic can be computed easily even when many independent variables are involved.

In the general model with k explanatory variables and an intercept, suppose we have two groups, call them $g = 1$ and $g = 2$. We would like to test whether the intercept and all slopes are the same across the two groups. Write the model as

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + \ldots + \beta_{g,k}x_k + u, \tag{7.23}$$

for $g = 1$ and $g = 2$. The hypothesis that each beta in (7.23) is the same across the two groups involves $k + 1$ restrictions (in the GPA example, $k + 1 = 4$). The unrestricted model, which we can think of as having a group dummy variable and k interaction terms

in addition to the intercept and variables themselves, has $n - 2(k + 1)$ degrees of freedom. [In the GPA example, $n - 2(k + 1) = 366 - 2(4) = 358$.] So far, there is nothing new. The key insight is that the sum of squared residuals from the unrestricted model can be obtained from two *separate* regressions, one for each group. Let SSR_1 be the sum of squared residuals obtained estimating (7.23) for the first group; this involves n_1 observations. Let SSR_2 be the sum of squared residuals obtained from estimating the model using the second group (n_2 observations). In the previous example, if group 1 is females, then $n_1 = 90$ and $n_2 = 276$. Now, the sum of squared residuals for the unrestricted model is simply $SSR_{ur} = SSR_1 + SSR_2$. The restricted sum of squared residuals is just the SSR from pooling the groups and estimating a single equation, say SSR_P. Once we have these, we compute the F statistic as usual:

$$F = \frac{[SSR_P - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k + 1)]}{k + 1}, \tag{7.24}$$

where n is the *total* number of observations. This particular F statistic is usually called the **Chow statistic** in econometrics. Because the Chow test is just an F test, it is only valid under homoskedasticity. In particular, under the null hypothesis, the error variances for the two groups must be equal. As usual, normality is not needed for asymptotic analysis.

To apply the Chow statistic to the GPA example, we need the SSR from the regression that pooled the groups together: this is $SSR_P = 85.515$. The SSR for the 90 women in the sample is $SSR_1 = 19.603$, and the SSR for the men is $SSR_2 = 58.752$. Thus, $SSR_{ur} = 19.603 + 58.752 = 78.355$. The F statistic is $[(85.515 - 78.355)/78.355](358/4) \approx 8.18$; of course, subject to rounding error, this is what we get using the R-squared form of the test in the models with and without the interaction terms. (A word of caution: there is no simple R-squared form of the test if separate regressions have been estimated for each group; the R-squared form of the test can be used only if interactions have been included to create the unrestricted model.)

One important limitation of the Chow test, regardless of the method used to implement it, is that the null hypothesis allows for no differences at all between the groups. In many cases, it is more interesting to allow for an intercept difference between the groups and then to test for slope differences; we saw one example of this in the wage equation in Example 7.10. There are two ways to allow the intercepts to differ under the null hypothesis. One is to include the group dummy and all interaction terms, as in equation (7.22), but then test joint significance of the interaction terms only. The second is to form an F statistic as in equation (7.24), but where the restricted sum of squares, called "SSR_P" in equation (7.24), is obtained by the regression that allows an intercept shift only. In other words, we run a pooled regression and just include the dummy variable that distinguishes the two groups. In the grade point average example, we regress *cumgpa* on *female*, *sat*, *hsperc*, and *tothrs* using the data for male and female student-athletes. In the GPA example, we use the first method, and so the null is H_0: $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = 0$ in equation (7.20). (δ_0 is not restricted under the null.) The F statistic for these three restrictions is about 1.53, which gives a p-value equal to .205. Thus, we do not reject the null hypothesis.

Failure to reject the hypothesis that the parameters multiplying the interaction terms are all zero suggests that the best model allows for an intercept difference only:

$$\widehat{cumgpa} = 1.39 + .310\,female + .0012\,sat - .0084\,hsperc + .0025\,tothrs \quad (7.25)$$
$$(.18)\ (.059)\qquad (.0002)\qquad (.0012)\qquad (.0007)$$
$$n = 366,\ R^2 = .398,\ \bar{R}^2 = .392.$$

The slope coefficients in (7.25) are close to those for the base group (males) in (7.22); dropping the interactions changes very little. However, *female* in (7.25) is highly significant: its t statistic is over 5, and the estimate implies that, at given levels of *sat*, *hsperc*, and *tothrs*, a female athlete has a predicted GPA that is .31 point higher than that of a male athlete. This is a practically important difference.

7.5 A Binary Dependent Variable: The Linear Probability Model

By now, we have learned much about the properties and applicability of the multiple linear regression model. In the last several sections, we studied how, through the use of binary independent variables, we can incorporate qualitative information as explanatory variables in a multiple regression model. In all of the models up until now, the dependent variable y has had *quantitative* meaning (for example, y is a dollar amount, a test score, a percentage, or the logs of these). What happens if we want to use multiple regression to *explain* a qualitative event?

In the simplest case, and one that often arises in practice, the event we would like to explain is a binary outcome. In other words, our dependent variable, y, takes on only two values: zero and one. For example, y can be defined to indicate whether an adult has a high school education; y can indicate whether a college student used illegal drugs during a given school year; or y can indicate whether a firm was taken over by another firm during a given year. In each of these examples, we can let $y = 1$ denote one of the outcomes and $y = 0$ the other outcome.

What does it mean to write down a multiple regression model, such as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \quad (7.26)$$

when y is a binary variable? Because y can take on only two values, β_j cannot be interpreted as the change in y given a one-unit increase in x_j, holding all other factors fixed; y either changes from zero to one or from one to zero (or does not change). Nevertheless, the β_j still have useful interpretations. If we assume that the zero conditional mean assumption MLR.4 holds, that is, $E(u|x_1,\ldots,x_k) = 0$, then we have, as always,

$$E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

where $\mathbf{x}$ is shorthand for all of the explanatory variables.

The Chow Test for Structural Change across Time

In Chapter 7, we discussed how the Chow test—which is simply an F test—can be used to determine whether a multiple regression function differs across two groups. We can apply that test to two different time periods as well. One form of the test obtains the sum of squared residuals from the pooled estimation as the restricted SSR. The unrestricted SSR is the sum of the SSRs for the two separately estimated time periods. The mechanics of computing the statistic are exactly as they were in Section 7.4. A heteroskedasticity-robust version is also available (see Section 8.2).

Example 13.2 suggests another way to compute the Chow test for two time periods by interacting each variable with a year dummy for one of the two years and testing for joint significance of the year dummy and all of the interaction terms. Since the intercept in a regression model often changes over time (due to, say, inflation in the housing price example), this full-blown Chow test can detect such changes. It is usually more interesting to allow for an intercept difference and then to test whether certain slope coefficients changed over time (as we did in Example 13.2).

A Chow test can also be computed for more than two time periods. Just as in the two-period case, it is usually more interesting to allow the intercepts to change over time and then test whether the slope coefficients have changed over time. We can test the constancy of slope coefficients generally by interacting all of the time period dummies (except that defining the base group) with one, several, or all of the explanatory variables and test the joint significance of the interaction terms. Computer Exercises C13.1 and C13.2 are examples. For many time periods and explanatory variables, constructing a full set of interactions can be tedious. Alternatively, we can adapt the approach described in part (vi) of Computer Exercise C7.11. First, estimate the restricted model by doing a pooled regression allowing for different time intercepts; this gives SSR_r. Then, run a regression for each of the, say, T time periods and obtain the sum of squared residuals for each time period. The unrestricted sum of squared residuals is obtained as $SSR_{ur} = SSR_1 + SSR_2 + ... + SSR_T$. If there are k explanatory variables (not including the intercept or the time dummies) with T time periods, then we are testing $(T - 1)k$ restrictions and there are $T + Tk$ parameters estimated in the unrestricted model. So, if $n = n_1 + n_2 + ... + n_T$ is the total number of observations, then the df of the F test are $(T - 1)k$ and $n - T - Tk$. We compute the F statistic as usual: $[(SSR_r - SSR_{ur})/SSR_{ur}][(n - T - Tk)/(T - 1)k]$. Unfortunately, as with any F test based on sums of squared residuals or R-squareds, this test is not robust to heteroskedasticity (including changing variances across time). To obtain a heteroskedasticity-robust test, we must construct the interaction terms and do a pooled regression.

13.2 Policy Analysis with Pooled Cross Sections

Pooled cross sections can be very useful for evaluating the impact of a certain event or policy. The following example of an event study shows how two cross-sectional data sets, collected before and after the occurrence of an event, can be used to determine the effect on economic outcomes.

15

Instrumental Variables Estimation and Two Stage Least Squares

In this chapter, we further study the problem of **endogenous explanatory variables** in multiple regression models. In Chapter 3, we derived the bias in the OLS estimators when an important variable is omitted; in Chapter 5, we showed that OLS is generally inconsistent under **omitted variables**. Chapter 9 demonstrated that omitted variables bias can be eliminated (or at least mitigated) when a suitable proxy variable is given for an unobserved explanatory variable. Unfortunately, suitable proxy variables are not always available.

In the previous two chapters, we explained how fixed effects estimation or first differencing can be used with panel data to estimate the effects of time-varying independent variables in the presence of *time-constant* omitted variables. Although such methods are very useful, we do not always have access to panel data. Even if we can obtain panel data, it does us little good if we are interested in the effect of a variable that does not change over time: first differencing or fixed effects estimation eliminates time-constant explanatory variables. In addition, the panel data methods that we have studied so far do not solve the problem of time-varying omitted variables that are correlated with the explanatory variables.

In this chapter, we take a different approach to the endogeneity problem. You will see how the method of instrumental variables (IV) can be used to solve the problem of endogeneity of one or more explanatory variables. The method of two stage least squares (2SLS or TSLS) is second in popularity only to ordinary least squares for estimating linear equations in applied econometrics.

We begin by showing how IV methods can be used to obtain consistent estimators in the presence of omitted variables. IV can also be used to solve the **errors-in-variables** problem, at least under certain assumptions. The next chapter will demonstrate how to estimate simultaneous equations models using IV methods.

Our treatment of instrumental variables estimation closely follows our development of ordinary least squares in Part 1, where we assumed that we had a random sample from the underlying population. This is a desirable starting point because, in addition to simplifying the notation, it emphasizes that the important assumptions for IV estimation are stated in terms of the underlying population (just as with OLS). As we showed in Part 2, OLS can be applied to time series data, and the same is true of instrumental variables methods. Section 15.7 discusses some special issues that arise when IV methods are applied to time series data. In Section 15.8, we cover applications to pooled cross sections and panel data